Protein Science (2000), *9*:2278–2284. Cambridge University Press. Printed in the USA.

FOR THE RECORD

Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments

IDDO FRIEDBERG, TOMMY KAPLAN, AND HANAH MARGALIT

Department of Molecular Genetics and Biotechnology, The Hebrew University, Hadassah Medical School,
POB 12272, Jerusalem 91120, Israel

(RECEIVED June 12, 2000; FINAL REVISION August 2, 2000; ACCEPTED August 24, 2000)

Abstract: The PSI-BLAST algorithm has been acknowledged as one of the most powerful tools for detecting remote evolutionary relationships by sequence considerations only. This has been demonstrated by its ability to recognize remote structural homologues and by the greatest coverage it enables in annotation of a complete genome. Although recognizing the correct fold of a sequence is of major importance, the accuracy of the alignment is crucial for the success of modeling one sequence by the structure of its remote homologue. Here we assess the accuracy of PSI-BLAST alignments on a stringent database of 123 structurally similar, sequence-dissimilar pairs of proteins, by comparing them to the alignments defined on a structural basis. Each protein sequence is compared to a nonredundant database of the protein sequences by PSI-BLAST. Whenever a pair member detects its pair-mate, the positions that are aligned both in the sequential and structural alignments are determined, and the alignment sensitivity is expressed as the percentage of these positions out of the structural alignment. Fifty-two sequences detected their pair-mates (for 16 pairs the success was bi-directional when either pair member was used as a query). The average percentage of correctly aligned residues per structural alignment was 43.5 ± 2.2%. Other properties of the alignments were also examined, such as the sensitivity vs. specificity and the change in these parameters over consecutive iterations. Notably, there is an improvement in alignment sensitivity over consecutive iterations, reaching an average of 50.9 ± 2.5% within the five iterations tested in the current study.

Keywords: alignment accuracy; PSI-BLAST; sequence alignment; structure alignment

With the fast accumulation of solved structures in the structural database PDB, it has been become evident that many of the newly solved structures are classified into already known fold families. This is obvious for proteins that show significant sequence similarity to proteins of known structure, as it was shown that even 30% of sequence identity between two proteins may suffice to ensure their folding into similar structures (Sander & Schneider, 1991). However, it is most remarkable that there are many proteins with no discernible sequence similarity that adopt the same fold (e.g., Chothia, 1992; Orengo et al., 1994; Rost, 1999; Brenner & Levitt, 2000; Koppensteiner et al., 2000). Recent surveys of the PDB have reported that among the newly determined structures that show no sequence similarity to PDB sequences, about 75% fall into the same fold families of already known structures (Brenner & Levitt, 2000; Koppensteiner et al., 2000). Had we known to recognize these relationships from the protein sequences, the fold of many proteins could be identified without tedious structural experiments, which in turn could then focus on proteins with high probability to have a new fold (Kim, 1998).

Indeed, during the last 10 years new and more sensitive approaches for fold recognition have been developed that succeed to varying degrees in identifying the relationships between remote structural homologues. These include: (1) various threading approaches that evaluate the compatibility between a protein sequence and a given structural template (e.g., Jones et al., 1992; Sippl & Weitckus, 1992; Bryant & Lawrence, 1993; Fischer & Eisenberg, 1996; Rost et al., 1997). (2) More advanced sequence comparison procedures that take into account multiple alignments rather than pairwise comparisons only (Krogh et al., 1994; Altschul et al., 1997; Park et al., 1998; Rychlewski et al., 2000; Teichmann et al., 2000). Such approaches were shown to recognize three times as many remote homologues as pairwise methods (Park et al., 1998). Also, in several cases these methods were shown to outperform the threading methods, as demonstrated in the recent CASP3 meeting (for review, see Sternberg et al., 1999). (3) Recently, new approaches have been reported that incorporate both multiple sequence alignments and threading according to energy considerations, resulting in improved recognition of remote homologues (e.g., Jones, 1999; Panchenko et al., 2000).

Reprint requests to: Hanah Margalit, Department of Molecular Genetics and Biotechnology, The Hebrew University, Hadassah Medical School, POB 12272, Jerusalem 91120, Israel; e-mail: hanah@md2.huji.ac.il.

Abbreviations: PDB, Protein Data Bank; PSSM, position specific scoring matrix; PSI-BLAST, position specific iterated basic local alignment search tool; CASP, critical assessment of structure prediction; SSAP, sequential structure alignment program; FSSP, fold classification based on structure–structure alignments of proteins.

One of the most widely used algorithms is the PSI-BLAST, which belongs to the second category (Altschul et al., 1997). PSI-BLAST identifies remote homologues iteratively. When comparing a query sequence against the protein database, PSI-BLAST generates a position specific scoring matrix (PSSM) from the multiple alignment of sequences identified in that search. The PSSM is used to search the databases for additional similar sequences, and these are now used to update the matrix. This process is repeated until a predetermined number of iterations is reached, or until the searches converge (no new sequences are added). Using enough iterations, with a large enough database, weak but biologically meaningful similarities may be detected. PSI-BLAST has been acknowledged as one of the most powerful tools for detecting remote structural homologues using sequence considerations only (Park et al., 1998; Salamov et al., 1999). It has been extensively used in genome annotation and was shown to improve significantly the coverage obtained by pairwise procedures and to provide structural assignments to 19–47% of the sequences of different genomes (Muller et al., 1999; for review, see Teichmann et al., 1999).

Recognizing the fold family of a protein opens the possibility to model the structure of that protein based on the structure of its identified homologue. However, to achieve reliable models, the quality of the alignment itself is very important, as it would serve as the basis for homology modeling. Here we assess the alignment accuracy obtained by PSI-BLAST for remote structural homologues by comparing the sequence alignment to the structural alignment. For this we use a stringent database of 123 protein pairs that are structurally similar but sequentially dissimilar. The main question that we address is what percentage of residues in the structural alignment are correctly aligned in the sequence alignment obtained by PSI-BLAST, or, in other words, what is the sensitivity of the PSI-BLAST alignments. In addition, the process of analysis and the results enable us to answer several more questions: (1) how many sequences detect their pair-mates? (2) In what iteration number did the detection occur? (3) Is there a change in the alignment sensitivity with subsequent iterations? (4) What is the specificity of the alignment (fraction of correctly aligned residues out of the sequence alignment)? (5) How does the specificity change with subsequent iterations? (6) How are the correctly aligned residues distributed in respect to the secondary structure elements of the structures?

Results and discussion: The assessment was carried out on a database of pairs of remote structural homologues, constructed by rather stringent criteria. Two databases were used as a starting point: The fold classification based on structure–structure alignments of proteins (FSSP) (Holm & Sander, 1996) and Distant Aligned Protein Sequences (DAPS) (1998; Rice & Eisenberg, http://siren.bio.indiana.edu/daps). The DAPS database is based on FSSP and contains alignments of all protein pairs sharing less than 25% identical residues. These pairs of proteins were based on the PDB_SELECT25 list (Hobohm & Sander, 1994). For generation of our database, the DAPS database was filtered using the following criteria: (1) minimal protein length of 30 residues for both pair members, (2) resolution better than 3.5 Å for each pair member, (3) difference in sequence lengths within a protein pair does not exceed 50% of the shorter member, (4) the structural alignment length is at least 60% of the longer member's length, and (5) protein pairs whose similarity could be detected by the Smith–Waterman algorithm (Smith & Waterman, 1981) were excluded. At the end of this procedure, the database contained 123 pairs of structurally similar proteins with 12% average sequence identity (http://bioinfo.md.huji.ac.il/marg/SSSD).

Each chain in the database was submitted as a query to PSI-BLAST and run against the NR database (a nonredundant compilation of all known protein sequences). The program was run with default parameters for five iterations, or until convergence. The run under these conditions had several reasons: (1) We wished to perform our assessments based on the commonly used PSI-BLAST runs. (2) As our goal was not to assess the power of PSI-BLAST in fold recognition [this was studied extensively by Park et al. (1998) and Salamov et al. (1999)], but rather to assess the accuracy of the alignments, we did not try different parameters. (3) Park et al. (1998) demonstrated that although they allowed 20 iterations for PSI-BLAST, 61% of the queries finished their search after two to four iterations, and 18% after 5–10 iterations. Moreover, 39% of the queries found their homologues after four iterations. Again, as our goal was not to detect as many homologues as possible, but to assess the alignment accuracy of the detected ones, we set the limit arbitrarily at five iterations (six rounds). When a pair member detected its partner, we compared the alignment provided by PSI-BLAST to the structural alignment as reported in FSSP. The alignment accuracy was assessed by its sensitivity and specificity measures. The sensitivity of an alignment was evaluated by determining the number of aligned positions that existed both in the sequence alignment and in the structural alignment, and dividing this number by the number of aligned positions in the structural alignment. This ratio is expressed as a percentage (×100). The specificity of the alignment was calculated as the number of the correctly aligned residues divided by the number of residues in the sequence alignment (×100). Such measures were used previously to assess the accuracy of alignments obtained by fold recognition methods (e.g., Marchler-Bauer & Bryant, 1997; Rost et al., 1997; Russell et al., 1998; Jones, 1999; Domingues et al., 2000).

The results of the analysis are summarized in Table 1. Fifty-two of the queries detected their partner sequence (for 16 pairs detection occurred when the search was carried out with either of the pair members as a query). This means that PSI-BLAST succeeded to detect the remote structural homologues in 21% of 246 searches, or, in other words, for 36 of the 123 pairs their pair-mate was detected (29.3%). Because the sequence pairs in our database shared only 12% identical residues on the average, this rate of detection is comparable to that reported by Park et al. (1998) and Salamov et al. (1999), using less restrictive databases. Forty-eight percent of the detections occurred already at the first iteration (Fig. 1). The distribution of alignment sensitivities at the iterations where detections have occurred is demonstrated in Figure 2. As shown, 18 of the 52 alignments reached alignment sensitivity higher than 50% already at the detection iteration. The mean sensitivity was 43.5 ± 2.2% (standard error). Notably, in a recent paper that was published while this manuscript was under review, Sauder et al. (2000) also reported that PSI-BLAST correctly aligns 40% of the residues when the sequence identity of the protein pairs was 10–15%. In our study, in total, out of 11,062 structurally aligned residues, 4,683 were correctly aligned by PSI-BLAST (42.3%). The average specificity of the detection iteration was 56.6 ± 2.14%. A positive correlation was observed between the sensitivity and the specificity ($r = 0.51$ by Spearman rank correlation). There was no correlation between the sensitivity and the structural alignment's length or with the number of aligned residues in the structural

Table 1. *Alignment attributes of deteced pair-mates*[a]

		Detection iteration				Maximum sensitivity iteration			
Query	Target	Iteration No.	E-value	Sensitivity	Specificity	Iteration no.	E-value	Sensitivity	Specificity
1bbpA	1hbq	1	8.0E−07	61.8	74.6	5	2.0E−32	74.3	80.7
1ceo	1eceA	1	8.0E−16	60.2	58.0	3	8.0E−39	62.5	60.7
1cnv	2ebn[b]	1	2.0E+00	21.7	58.1	2	9.1E−01	22.7	66.2
1dbqA	1gca	1	1.0E−09	63.3	78.3	5	1.0E−35	87.9	86.5
1dhr	1xel[b]	1	5.8E−02	23.4	51.7	1	5.8E−02	23.4	51.7
1eceA	1ceo	1	1.0E−12	45.1	70.8	5	2.0E−34	58.3	54.0
1eceA	1edg	1	1.0E−18	40.1	40.2	5	3.0E−45	42.6	39.5
1edg	1eceA	1	1.0E−11	41.2	41.8	4	3.0E−41	44.8	39.6
1eny	1cydA	1	5.0E−36	64.7	72.3	3	5.0E−47	79.0	81.0
1fnc	1ndh	1	2.0E−14	41.9	55.2	5	4.0E−56	60.3	56.8
1fnc	2pia	1	9.0E−04	44.9	56.4	3	6.0E−23	67.6	69.5
1gal	3cox	1	2.0E−09	35.7	33.7	5	2.0E−83	45.9	40.2
1gca	1dbqA	1	4.0E−11	70.7	82.3	2	1.0E−59	87.5	85.5
1hbq	1bbpA	1	8.2E−02	49.3	64.1	5	4.0E−23	59.9	62.8
1mup	1bbpA	1	5.0E−04	58.2	71.3	5	4.0E−20	64.5	70.0
1mup	1hbq	1	2.0E−06	46.4	52.9	5	4.0E−31	68.6	62.3
1ndh	1fnc[c]	1	7.0E−22	49.8	57.9	4	6.0E−44	66.4	63.3
1pot	1sbp	1	1.0E−04	24.9	37.3	5	3.0E−35	34.0	34.0
1ptvA	1ytw	1	6.0E−19	52.4	62.3	5	5.0E−23	54.1	64.3
1vhrA	1ptvA	1	1.1E−01	16.0	100.0	5	5.0E−23	21.5	24.5
2gdm	3sdhA	1	6.0E−07	68.2	67.7	2	8.0E−25	71.9	70.8
2mnr	4enl	1	4.0E−03	51.6	54.2	4	3.0E−90	61.1	56.5
3cox	1gal	1	4.0E−03	9.5	74.5	5	7.0E−53	52.1	45.4
4enl	2mnr	1	2.4E+00	26.0	80.4	2	8.3E−01	26.0	80.4
1ctn	1cnv	2	6.8E+00	20.4	46.6	3	6.8E+00	26.8	43.2
1dhr	1cydA	2	3.0E−39	63.2	59.5	4	1.0E−43	65.6	64.6
1dhr	1eny	2	3.0E−13	61.1	63.2	3	6.0E−17	69.4	68.5
1ecpA	1pbn	2	1.0E−08	40.5	51.2	5	1.0E−33	48.2	56.1
1hbq	1mup	2	2.3E−02	51.4	59.5	5	3.0E−34	56.4	50.6
1opbA	1hbq	2	2.0E−04	49.1	62.8	5	4.0E−22	54.6	41.7
1pvc1	1bbt1	2	3.0E−04	40.4	65.1	5	6.0E−50	54.2	48.7
1sbp	1pot	2	4.0E−03	45.2	49.0	4	5.0E−37	45.5	48.0
1vic	2admA	2	2.3E−02	41.1	38.7	5	4.0E−16	42.5	39.7
1ytw	1ptvA	2	2.0E−63	53.7	61.7	5	4.0E−69	61.8	61.0
2ebn	1ctn	2	1.4E−02	28.8	54.9	5	2.0E−52	29.6	26.6
2por	2omf[b]	2	9.0E−36	27.1	25.9	3	1.0E−52	29.0	27.5
1ash	3sdhA[b]	3	8.0E−16	70.6	79.3	3	8.0E−16	70.6	79.3
1bbpA	1mup	3	5.0E−11	52.5	54.4	5	3.0E−31	54.6	55.0
1ceo	1xyzA	3	3.0E−07	30.3	49.7	4	3.0E−15	32.5	46.1
1tca	1broA	3	9.0E−05	17.0	43.4	5	4.0E−35	41.3	42.7
2ebn	1nar[b]	3	2.0E−03	28.9	48.9	3	2.0E−03	28.9	48.9
1531	1lzr	4	8.0E−01	20.6	50.0	5	9.2E−02	26.8	56.5
1hbq	1obpA	4	3.0E−03	40.9	32.4	5	3.0E−08	50.9	39.2
1httA	1sesA[b]	4	2.0E−20	53.6	51.1	4	2.0E−20	53.6	51.1
1pea	2dri	4	7.0E−05	38.0	41.5	5	1.0E−10	38.8	43.8
1vid	1xvaA	4	7.0E−16	45.8	57.0	5	8.0E−17	49.4	50.9
1bbt1	1pvc1	5	3.0E−35	47.6	49.1	5	3.0E−35	47.6	49.1
1tca	1ede	5	1.0E−12	32.7	56.4	5	1.0E−12	32.7	56.4
2ebn	1cnv	5	2.0E−24	27.8	23.8	5	2.0E−24	27.8	23.8
2pia	1fnc	5	6.0E−37	64.4	65.0	5	6.0E−37	64.4	65.0

[a] The list is sorted by the PSI-BLAST iteration in which the query sequence detected the target sequence. Sensitivity: percentage of correctly aligned positions by PSI-BLAST out of the structural alignment. Specificity: percentage of correctly aligned positions by PSI-BLAST out of the sequence alignment. Detection iteration: Iteration in which the target pair-mate was first detected. Maximum sensitivity iteration: iteration where maximum sensitivity was reached within the five iterations tested. E-values of the detection and maximum sensitivity iterations are given.

[b] Pair-mates that were initially detected and then lost at subsequent iterations.

[c] Converged after three iterations.

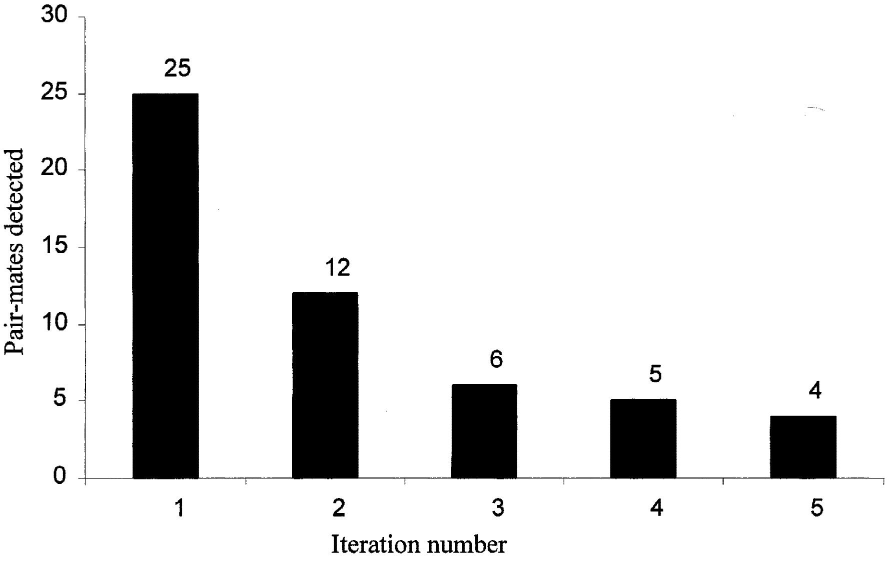

Fig. 1. Iteration number where initial detection of pair-mates occurred. The majority of pair-mates were detected at the first iteration.

alignment. We also examined whether there was a higher presence of the correctly aligned residues within secondary structure elements and found that in that respect they were randomly distributed along the structures; 83.2% of the correctly aligned residues were located within secondary structures, while the fraction of residues within secondary structures in the whole database was 81.7%.

Because each PSI-BLAST submission was run for five iterations, even in cases where the detection of the pair-mate occurred already in earlier iterations, we could follow the accuracy of the alignment as the iterations proceeded. We compared the alignment sensitivity between the iteration where detection has occurred and the subsequent iteration for all 52 pairs, and found a significant improvement ($p < 0.0005$ using the Wilcoxon paired samples test). The specificity, on the other hand, has not changed significantly. Interestingly, some of the detections disappeared in consecutive iterations (see Table 1). These phenomena result from the type of sequences accumulated as the PSSM is developed with successive iterations. It may happen that the new added sequences will drive the matrix to a different direction and cause the disappearance of already detected sequences, or that sequences that are closer to the detected sequence are added, leading to an improvement in the alignment. Overall, it seems that it is worthwhile to continue for several iterations to obtain better alignments, with higher sensitivity and no significant negative effect on the specificity. When we analyzed the best results for each of the 52 pairs (the iteration that yielded the highest alignment sensitivity, see Table 1 and Figs. 2 and 3), the average sensitivity was 50.9 ± 2.5%

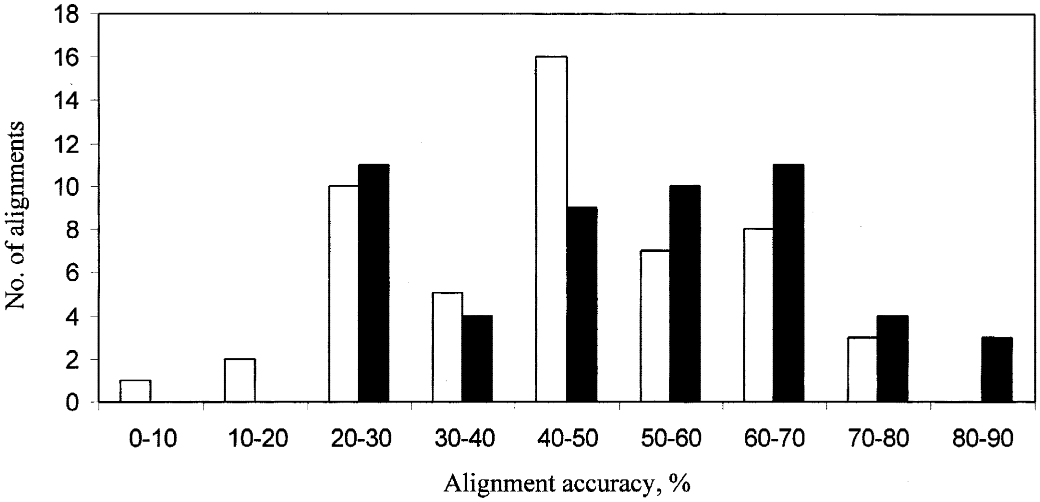

Fig. 2. Distribution of alignment sensitivities at the iteration where detection of pair-mates occurred (white bars) and at the maximum sensitivity iteration (black bars). See legend to Table 1 for definitions of sensitivity and of maximum sensitivity iteration.

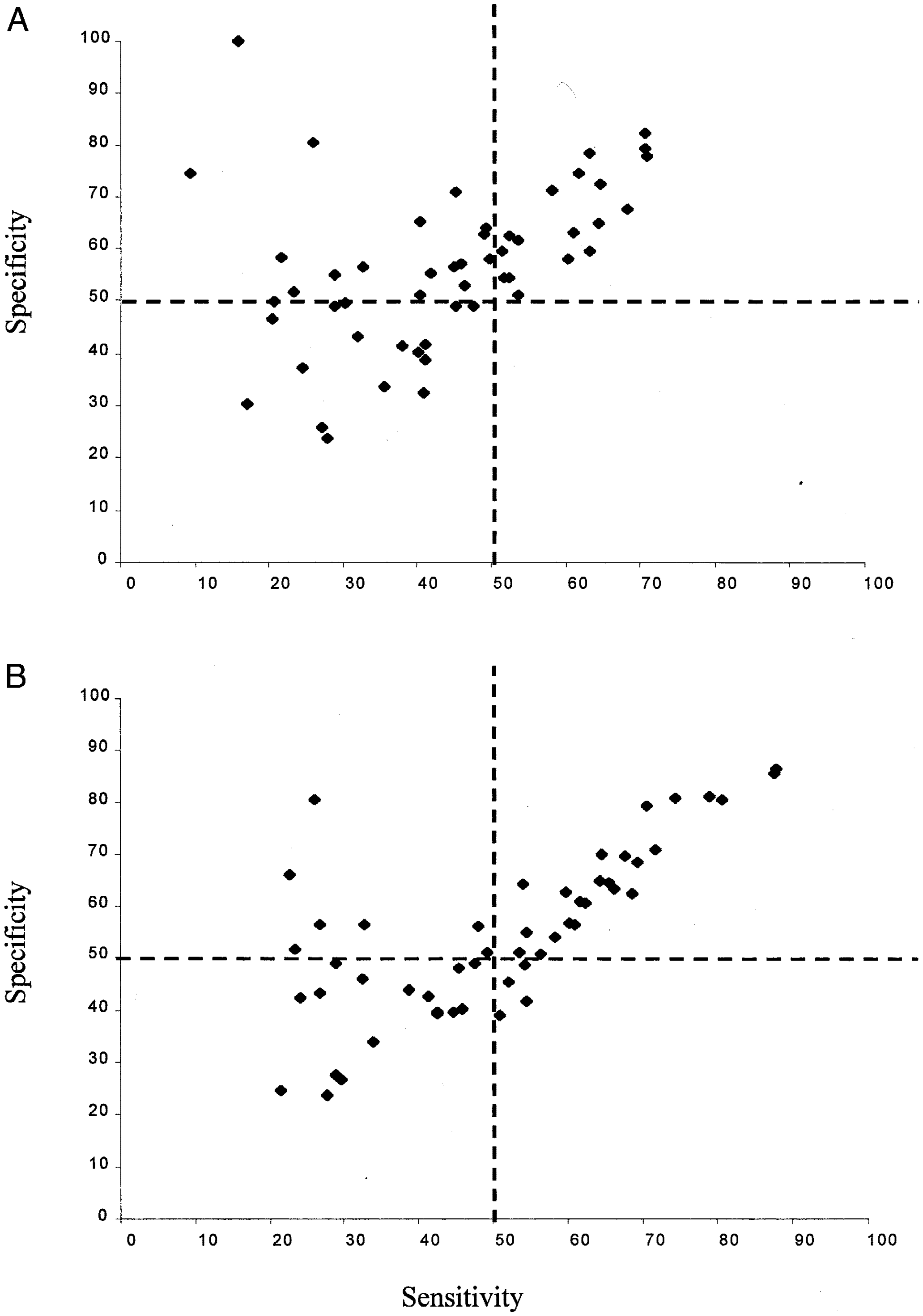

Fig. 3. Specificity vs. sensitivity: (**A**) at the detection iteration and (**B**) at the maximum sensitivity iteration (see legend to Table 1 for definition of maximum sensitivity iteration).

(with average specificity of 54.9 ± 2.1%). The detailed results may be viewed at http://bioinfo.md.huji.ac.il/marg/SSSD.

Notably, in 20 pairs the detection occurred only for one of the pair-members used as a query. This may mean that the PSSM developed when this pair member was used as a query has advanced in the direction of its pair-mate, while the other matrix (derived from the pair-mate as a query) has drifted away. It is interesting also to compare the bi-directional detections, when both pair members were used as queries and detected one another. In most cases, the detection occurred at the same iteration number and the alignment sensitivities were very similar. However, different scenarios were also observed. There were cases where the detection in one direction occurred at the first iteration, and in the other direction at the fifth iteration, but the alignments sensitivities were very similar (e.g., 2pia, 1fnc; see Table 1). In other cases, the detections may have occurred at the same iteration, but their sensitivities differed significantly (e.g., 2mnr, 4enl; see Table 1).

Our evaluation of alignment accuracy relies on the structural alignments reported in FSSP, based on the DALI algorithm for structural alignment (Holm & Sander, 1993). Evidently, different algorithms for structure alignment may provide different alignments (Godzik, 1996), resulting in different listings of the structurally aligned positions. In previous studies when such assessments were performed to evaluate the alignments obtained by fold recognition procedures (Marchler-Bauer & Bryant, 1997; Rost et al., 1997; Jones, 1999; Domingues et al., 2000) and by pairwise alignments (Domingues et al., 2000), various approaches were used. Rost et al. (1997) used the FSSP based on DALI for the structural alignments, while Jones (1999) used the alignments based on the SSAP algorithm (Orengo et al., 1992). We compared the DALI alignments of the pairs of proteins in our database to the alignments obtained by SSAP and found them to be very similar. This was done by the same approach that was used to assess the PSI-BLAST alignments (determination of co-aligned pairs of residues in both alignments). We found that ~80% of the paired positions were co-aligned when either alignment was used as a standard. Similarly, Sauder et al. (2000) compared the alignments in FSSP to structural alignments obtained by the combinatorial extension structure alignment program (Shindyalov & Bourne, 1998) and reported that 75% of the paired positions were co-aligned. Thus, to a very large degree the FSSP alignments are consistent with the alignments obtained by other algorithms. Marchler-Bauer & Bryant (1997) used a jury decision based on different structural alignment methods to determine the residues that were structurally aligned. In comparison to that approach, our assessment may only underestimate the sensitivity of the PSI-BLAST alignments, because our denominator should be either equal to or larger than that determined by taking a jury decision. Domingues et al. (2000) compared the alignments obtained by threading and pairwise sequence comparisons to structural alignments obtained by various procedures and reported their results based on the structural alignment to which the tested alignment fitted best. Taking that approach, our evaluation results could only be improved using other procedures for structural alignment. Taken together, we believe that our analysis provides a reliable evaluation of the accuracy of PSI-BLAST alignments of remote homologues.

With the introduction of the fold recognition procedures, a major concern regarded the disparity between the sequence-based and the structure-based alignments of remote homologues (Rost et al., 1997), as they should be used as a basis for homology modeling. In quite a few cases, while the correct fold was recognized, the alignment was rather poor when compared to the structural alignment. Two recent publications have reported that computational procedures that combine sequence information with knowledge-based pairwise potentials and solvent interactions predictions provide higher success in fold recognition. The reported alignments in those studies reach better than 50% sensitivity (Jones, 1999; Domingues et al., 2000). Jones (1999) reported an average alignment sensitivity of 46.2% for all cases when the correct fold was recognized. For about one-third of the pairs in his database, the alignment sensitivity exceeded 50%. Domingues et al. (2000) calculated the percentage of correctly aligned residues out of all aligned positions in their database and reported a sensitivity of 51%. Improvement in the alignment accuracy for folds recognized by such procedures has been also acknowledged in the recent CASP3 meeting (Marchler-Bauer & Bryant, 1999). Given the fact that PSI-BLAST uses sequence considerations only, its relative success in fold recognition is impressive. Moreover, in view of the stringent database used here of protein pairs with an average of 12% identical residues, the average alignment sensitivities obtained by PSI-BLAST of 43.5% in the detection iteration and of 50.9% in subsequent iterations are quite satisfactory.

Acknowledgments

We thank Nathan Linial for valuable discussions. This study was partially supported by the Israeli Science Foundation.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res 25*:3389–3402.

Brenner SE, Levitt M. 2000. Expectations from structural genomics. *Protein Sci 9*:197–200.

Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins 16*:92–112.

Chothia C. 1992. Proteins. One thousand families for the molecular biologist. *Nature 357*:543–544.

Domingues FS, Lackner P, Andreeva A, Sippl MJ. 2000. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J Mol Biol 297*:1003–1013.

Fischer D, Eisenberg D. 1996. Protein fold recognition using sequence-derived predictions. *Protein Sci 5*:947–955.

Godzik A. 1996. The structural alignment between two proteins: Is there a unique answer? *Protein Sci 5*:1325–1338.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Holm L, Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol 233*:123–138.

Holm L, Sander C. 1996. Mapping the protein universe. *Science 273*:595–603.

Jones DT. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol 287*:797–815.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.

Kim SH. 1998. Shining a light on structural genomics. *Nat Struct Biol 5 Suppl*:643–645.

Koppensteiner WA, Lackner P, Wiederstein M, Sippl MJ. 2000. Characterization of novel proteins based on known protein structures. *J Mol Biol 296*:1139–1152.

Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol 235*:1501–1531.

Marchler-Bauer A, Bryant SH. 1997. Measures of threading specificity and accuracy. *Proteins Suppl 1*:74–82.

Marchler-Bauer A, Bryant SH. 1999. A measure of progress in fold recognition? *Proteins Suppl 3*:218–225.

Muller A, MacCallum RM, Sternberg MJ. 1999. Benchmarking PSI-BLAST in genome annotation. *J Mol Biol 293*:1257–1271.

Orengo CA, Brown NP, Taylor WR. 1992. Fast structure alignment for protein databank searching. *Proteins Struct Funct Genet 14*:139–167.

Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature 372*:631–634.

Panchenko AR, Marchler-Bauer A, Bryant SH. 2000. Combination of threading potentials and sequence profiles improves fold recognition. *J Mol Biol 296*:1319–1331.

Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. 1998. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol 284*:1201–1210.

Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng 12*:85–94.

Rost B, Schneider R, Sander C. 1997. Protein fold recognition by prediction-based threading. *J Mol Biol 270*:471–480.

Russell RB, Saqi MA, Bates PA, Sayle RA, Sternberg MJ. 1998. Recognition of analogous and homologous protein folds—Assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng 11*:1–9.

Rychlewski L, Jaroszewski L, Li W, Godzik A. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci 9*:232–241.

Salamov AA, Suwa M, Orengo CA, Swindells MB. 1999. Genome analysis: Assigning protein coding regions to three-dimensional structure. *Protein Sci 8*:771–777.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins 9*:56–68.

Sauder JM, Arthur JW, Dunbrack RL. 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins 40*:6–22.

Shindyalov IN, Bourne PE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng 11*:739–747.

Sippl MJ, Weitckus S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins 13*:258–271.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol 147*:195–197.

Sternberg MJ, Bates PA, Kelley LA, MacCallum RM. 1999. Progress in protein structure prediction: Assessment of CASP3. *Curr Opin Struct Biol 9*:368–373.

Teichmann SA, Chothia C, Church GM, Park J. 2000. Fast assignments of protein structures to sequences using the intermediate sequence library. *Bioinformatics 16*:117–124.

Teichmann SA, Chothia C, Gerstein M. 1999. Advances in structural genomics. *Curr Opin Struct Biol 9*:390–399.